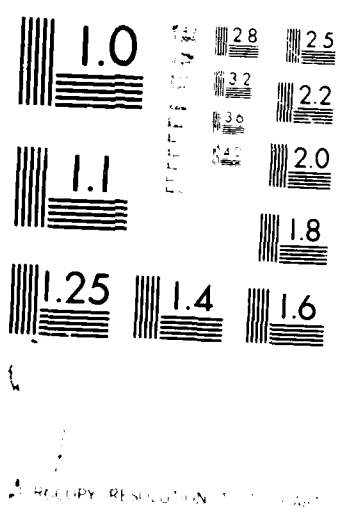
AD-A186 097 DESCRIPTIVE STATISTICS USING THE VETERANS 1/1
ADMINISTRATION FILE MANAGER AS A.. (U) NAVAL HEALTH
RESEARCH CENTER SAN DIEGO CA D R HODGINS 16 MAR 87
UNCLASSIFIED NHRC-87-22 F/G 12/5 NL


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0

Naval Health Research Center

DESCRIPTIVE STATISTICS USING THE VETERANS ADMINISTRATION FILE MANAGER AS A RELATIONAL DATABASE MANAGEMENT SYSTEM

D. R. HODGINS

REPORT NO. 87-22

NAVAL HEALTH RESEARCH CENTER
P.O. BOX 85122
SAN DIEGO, CALIFORNIA 92138

NAVAL MEDICAL RESEARCH AND DEVELOPMENT COMMAND
BETHESDA, MARYLAND

Descriptive Statistics Using the Veterans Administration File Manager as a Relational Database Management System

Dallas R. Hodgins

Medical Information Systems Department
Naval Health Research Center
P. O. Box 85122
San Diego, California 92138-9174

Report 87-22, supported by the Naval Medical Research and Development Command, Department of the Navy under work unit M0095.005-1053. The views expressed in this article are those of the author(s) and do not reflect the official policy or position of the Department of the Navy, Department of Defense, nor the U.S. Government. Approved for public release, distribution unlimited.

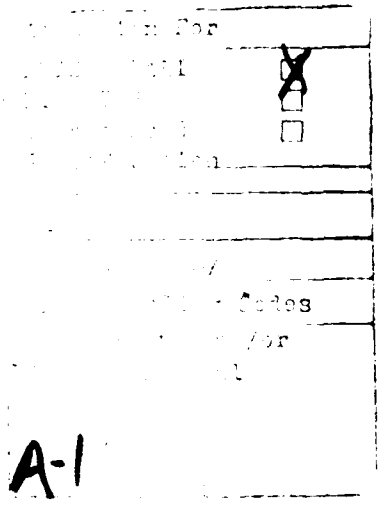
The author would like to acknowledge his colleague, Kathryn Medrano. Her patience and skills were indispensable in the preparation of this manuscript.

Executive Summary

The way data are stored in a computerized database determines ease of access and use. Using the Veterans Administration File Manager (FM) as a database management system allows the user to view the data as rows of records whose fields are the columns of a matrix. This "rectangular" structure is a relational database.

A simple four-variable database is developed while guiding the reader through the process of organizing the data using FM. The structure created is shown to have relational database properties so that the statistical routines that are implemented are seen to be logically related to the database.

The techniques illustrated support the view that using linear algebraic concepts enable the programmer to realize maximum efficiency in exploiting both the MUMPS language and the data being manipulated. Moreover, the end user, however unsophisticated, is able to form a clear picture of the data and can easily extract statistical information, create reports, and archive information that is meaningful and accessible.



Abstract

The development of descriptive statistics programs underscore the efficacy of a relational database. Relational arrays being amenable to algebraic manipulation allow efficient analysis and i dependent data bases. The Veteran's Administration File Manager is viewed as a relational database management system.

Introduction

The Naval Health Research Center (NHRC) developed the Navy Occupational Health Information Monitoring System (NOHIMS) to provide a method for satisfying the requirements of the Occupational Safety and Health Act of 1970. NOHIMS is presently being installed in major industrial centers operated by the Navy. These facilities typically employ from 6,000 to 10,000 individuals and, in turn, provide ancillary service to smaller units. There are two major components of NOHIMS. The medical component uses the Computer Stored Ambulatory Record System (COSTAR) for handling medical data, and the industrial component currently employs the Veteran's Administration File Manager (FM) for database management. The data gathered ensures that all individuals exposed to hazardous materials are identified, given periodic examinations, and their environment monitored, and will produce a solid basis for epidemiological studies[1].

The analysis, structuring, and storing of these data for present and future use by industrial hygienists, medical providers, and statisticians is evolving at NHRC in three phases: the development of descriptive statistics (presented herein); the development of derivative statistics such as various "t" tests, univariate and multiple regression analysis; and the linking of the NOHIMS database to the Statistics Program for Social Scientists (SPSS) package. This progression reflects an increasing statistical sophistication, as well as the practical limitations of handling different orders of magnitude of data. The programs presented in this paper are designed to handle the day-to-day demands of industrial hygienists dealing with sample sizes of a few hundred cases or less.

Using a relational model presents the data in its natural structure affording "a sound basis for treating derivability, redundancy, and consis-

tency of relations." Mathematically, a relation R is defined on n sets S_1, S_2,...S_n if R is a set of n-tuples such that each set has its first element from S_1, its second element from S_2, etc. where S_i is the i'th domain of R[2]. Since MUMPS globals can be viewed as "flat tables", it is not difficult to develop a data manipulation methodology to achieve data independence with "user friendly" algebraic database language commands[3]. The resolution of these issues is a prerequisite to structuring a useful archive; moreover, the fact that the data structure of FM can be viewed as a relational array is the salient point of this paper. With a "rectangular" view of the data, the use of linear algebra leads to simple, efficient algorithms as shown in the material presented.

Methods and Results

At present, NOHIMS is being implemented using Intersystems MUMPS M/VX running under VMS 4.3 on Digital Equipment 8200 machines. The following discussion conforms to these constraints and assumes FM version 17.0 or later. From the NOHIMS main menu, calling FM presents the FM menu:

CHOOSE FROM:

1 ENTER OR EDIT FILE ENTRIES
2 PRINT FILE ENTRIES
3 SEARCH FILE ENTRIES
4 MODIFY FILE ATTRIBUTES
5 INQUIRE TO FILE ENTRIES
6 UTILITY FUNCTIONS
7 STATISTICS
8 LIST FILE ATTRIBUTES
9 TRANSFER ENTRIES

from which we will choose, in order, options 4, 1, and 2 to create a file structure, enter data, and print the data to use in the statistical routines. The file will be called STATISTICS and will consist of 32 records with five fields in each record - NAME, AGE, EDUC, INCOME, and NC (number of children). The fields will be defined as numeric except for name--CASE1, CASE2, etc.--which is free text.

Creating a File

Choosing option 4, the dialogue with FM is straightforward. All user responses are underlined. Field position 1 is always assumed by FM to be the NAME field and has field number .01. Subsequent fields are numbered 1, 2, etc. This can create confusion if not fixed firmly in one's mind. As you can see, this makes the "^" piece position of AGE, our 1 field, the "^" piece position 2.

```
Select Option: 4 MODIFY FILE ATTRIBUTES
MODIFY WHAT FILE: STATISTICS
   ARE YOU ADDING STATISTICS AS A NEW FILE? Y (YES)
   FILE NUMBER: 1000//
INTERNAL GLOBAL REFERENCE:   ^DIZ(1000,//
Select FIELD: AGE
   ARE YOU ADDING 'AGE' AS A NEW FIELD? (THE 2ND)? Y (YES)
   FIELD NUMBER: 1//
DATA TYPE OF AGE: NUMERIC
INCLUSIVE LOWER BOUND: 0
INCLUSIVE UPPER BOUND: 100
IS THIS A DOLLAR AMOUNT (Y/N): NO//
MAXIMUM NUMBER OF FRACTIONAL DIGITS: 0//
WILL AGE FIELD BE MULTIPLE? NO// (NO)
SUBSCRIPT: 0//
"^"-PIECE POSITION: 2//
IS AGE ENTRY MANDATORY (Y/N): NO//
'HELP' PROMPT: TYPE A WHOLE NUMBER BETWEEN
     0 AND 100
```

Continuing in a like manner at each select field prompt, the attributes of EDUC, INCOME, and NC are entered, completing the definition of the file STATISTICS.

The results of this activity are stored in the data dictionary global ^DD and can be viewed directly by calling ^DD with the ^%G utility or by using the LIST FILE ATTRIBUTES option, entering "brief" after the double slashes following STANDARD and "NO" to alphabetizing. The result is:

```
DATA DICTIONARY #1000 -- STATISTICS FILE
NAME      1000,.01    FREE TEXT
AGE       1000,1      NO. BTWN 0 AND 100
EDUC      1000,2      NO. BTWN 0 AND 20
INCOME    1000,3      NO. BTWN 5000 AND 25000
NC        1000,4      NO. BTWN 1 AND 4
```

Entering Data

Now, to enter data in the file STATISTICS, choose FM option 1 and follow the prompts:

```
Select OPTION: 1 ENTER OR EDIT FILE ENTRIES
INPUT TO WHAT FILE:  STATISTICS//
EDIT WHICH FIELD:  ALL//
Select STATISTICS NAME: CASE1
ARE YOU ADDING 'CASE1' AS A NEW STATISTICS THE 1ST)? Y (YES)
AGE:  57
EDUC:  4
INCOME:  6281
NC:  3
```

Entering 31 more records results in the creation of the ^DIZ global:

```
^DIZ=DATA GLOBAL "STATISTICS (FILE 1000)"
^DIZ(1000,0)=STATISTICS^1000^32^32
^DIZ(1000,1,0)=CASE1^57^4^6281^3
^DIZ(1000,2,0)=CASE2^20^4^10516^2
^DIZ(1000,3,0)=CASE3^76^6^6898^3
^DIZ(1000,4,0)=CASE4^47^6^8212^3
^DIZ(1000,5,0)=CASE5^44^6^11744^2
^DIZ(1000,6,0)=CASE6^64^8^8618^3
^DIZ(1000,7,0)=CASE7^31^8^10011^2
^DIZ(1000,8,0)=CASE8^47^8^12405^2
^DIZ(1000,9,0)=CASE9^38^8^14664^1
^DIZ(1000,10,0)=CASE10^27^10^7472^4
```

```
^DIZ(1000,11,0)=CASE11^22^10^11598^2
^DIZ(1000,12,0)=CASE12^39^10^15366^3
^DIZ(1000,13,0)=CASE13^60^11^10186^2
^DIZ(1000,14,0)=CASE14^25^12^9771^3
^DIZ(1000,15,0)=CASE15^51^12^12444^4
^DIZ(1000,16,0)=CASE16^33^12^14213^3
^DIZ(1000,17,0)=CASE17^25^12^16908^2
^DIZ(1000,18,0)=CASE18^47^12^18347^3
^DIZ(1000,19,0)=CASE19^52^13^19546^1
^DIZ(1000,20,1)=CASE20^31^14^12660^1
^DIZ(1000,21,0)=CASE21^42^14^16326^3
^DIZ(1000,22,0)=CASE22^56^15^12772^1
^DIZ(1000,23,0)=CASE23^38^15^17218^3
^DIZ(1000,24,0)=CASE24^27^16^12599^2
^DIZ(1000,25,0)=CASE25^30^16^14852^3
^DIZ(1000,26,0)=CASE26^37^16^19138^4
^DIZ(1000,27,0)=CASE27^20^16^21779^2
^DIZ(1000,28,0)=CASE28^54^17^16428^4
^DIZ(1000,29,0)=CASE29^46^17^20018^4
^DIZ(1000,30,0)=CASE30^41^18^16526^4
^DIZ(1000,31,0)=CASE31^53^18^19414^4
^DIZ(1000,32,0)=CASE32^24^20^18822^2
```

which represents the mathematical relationship STATISTICS with domains NAME, AGE, EDUC, INCOME, and NC having values CASE1, 57, 4, 6281, and 3 respectively for set 1 of the 32 sets.

Print Entries

In order to exercise the statistic routines Histogram and Scattergram (which use the original FM routines ^DIH and ^DIG), the global ^DOSV must be created. To do so, there are various ploys involving the !, &, +, and # signs preceding the variable name in the Print File Entries option. For the sake of clarity and simplicity, what follows demonstrates what is needed (the Search File Entries option also is being slighted). The reader is referred to the VA FILEMAN USER'S MANUAL (1986) for further details[4].

Choosing Option 2, we respond as follows:

```
OUTPUT FROM WHAT FILE: STATISTICS//
SORT BY:  NAME// +EDUC
START WITH EDUC: FIRST//
  WITHIN EDUC, SORT BY: +INCOME
  START WITH INCOME: FIRST//
    WITHIN INCOME, SORT BY: NAME
    START WITH NAME: FIRST//
      WITHIN NAME, SORT BY:
STORE IN 'SORT' TEMPLATE:
FIRST PRINT FIELD:  #AGE
THEN PRINT FIELD:  #EDUC
THEN PRINT FIELD:  #INCOME
THEN PRINT FIELD:  #NC
THEN PRINT FIELD:
```

The results are stored in the global ^DOSV:

```
^DOSV(0,"TTB2:",0,1,"H")=76
                    "L")=20
                    "N")=32
                    "Q")=59268
                    "S")=1304
^DOSV(0,"TTB2:",0,2,"H")=20
                    "L")=4
                    "N")=32
                    "Q")=5198
                    "S")=384
^DOSV(0,"TTB2:",0,3,"H")=21779
                    "L")=6281
                    "N")=32
                    "Q")=6714653180
                    "S")=443752
^DOSV(0,"TTB2:",0,4,"H")=4
                    "L")=1
```

```
                    "N")=32
                    "Q")=255
                    "S")=85
^DOSV(0,"TTB2:",2,4,6281,1,"N")=1
                           "S")=57
^DOSV(0,"TTB2:",2,4,6281,2,"N")=1
                           "S")=4
^DOSV(0,"TTB2:",2,4,6281,3,"N")=1
                           "S")=6281
^DOSV(0,"TTB2:",2,4,6281,4,"N")=1
                           "S")=3
                .
                .
                .

^DOSV(0,"TTB2:",2,20,18822,1,"N")=1
                             "S")=24
^DOSV(0,"TTB2:",2,20,18822,2,"N")=1
                             "S")=20
^DOSV(0,"TTB2:",2,20,18822,3,"N")=1
                             "S")=18822
^DOSV(0,"TTB2:",2,20,18822,4,"N")=1
                             "S")=2
^DOSV(0,"TTB2:","BY",1)=1000^2^EDUC +
^DOSV(0,"TTB2:","F",1)=1000^1^AGE^NJ3,0
                    2)=1000^2^EDUC^NJ2,0
                    3)=1000^3^INCOME^NJ5,0
                    4)=1000^4^NC^NJ1,0
```

The main body of ^DOSV contains the same information inherent in ^DIZ but is obviously not as tractable. The first four entries have statistical information, and the last four list the file number, field position of the variable, the name of the variable, and its attributes (for "AGE" NJ3,0 means a three position right justified numeric field with no decimal digits).

Statistical Analysis

Now we are ready to calculate some statistics. Choosing the STATISTICS option of the FM menu and responding with a question mark to the first prompt, we get the statistical routine menu:

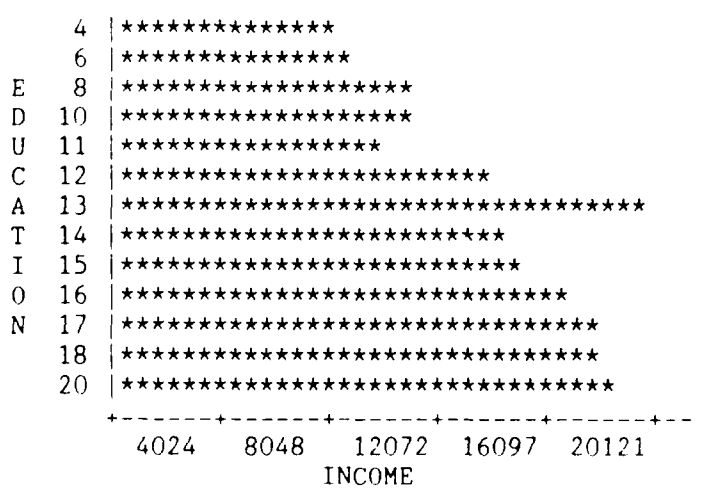
```
Select STATISTICAL ROUTINE: ?
CHOOSE FROM:
  1  DESCRIPTIVE STATISTICS
  2  SCATTERGRAM
  3  HISTOGRAM
  4  ESTIMATED LINEAR CORRELATION COEFFICIENTS
  5  COEFFICIENTS OF DETERMINATION
  6  RANDOM SAMPLE--DESCRIPTIVE STATISTICS
```

Choosing each item in turn produces:

1) Descriptive Statistics (see figure 1)

2) Scattergram (see figure 2)

3) Histogram of Income Versus Education:



I. MEASURES OF CENTRAL TENDENCY:

VARIABLE	N	MEAN	MEDIAN	MODE
AGE	32	40.750	40.000	47 OCCURS 3 TIMES
EDUC	32	12.000	12.000	12 OCCURS 5 TIMES
INCOME	32	13867.250	13492.500	NO MULTIPLE OCCURRENCES
NC	32	2.656	3.000	3 OCCURS 11 TIMES

II. MEASURES OF DISPERSION:

A. ACROSS THE RANGE OF DATA:

VARIABLE	N	MIMIMUM	MAXIMUM	25% TILE SCORE	75% TILE SCORE
AGE	32	20.000	76.000	27.00	51.00
EDUC	32	4.000	20.000	7.50	15.25
INCOME	32	6281.000	21779.000	10186.00	16908.00
NC	32	1.000	4.000	1.40	2.91

B. ABOUT THE MEAN:

VARIABLE	N	VARIANCE	STANDARD DEVIATION	STANDARD ERROR OF THE MEAN
AGE	32	197.742	14.062	2.486
EDUC	32	19.032	4.363	0.771
INCOME	32	18097847.032	4254.156	752.036
NC	32	0.943	0.971	0.172

Figure 1. Descriptive Statistics

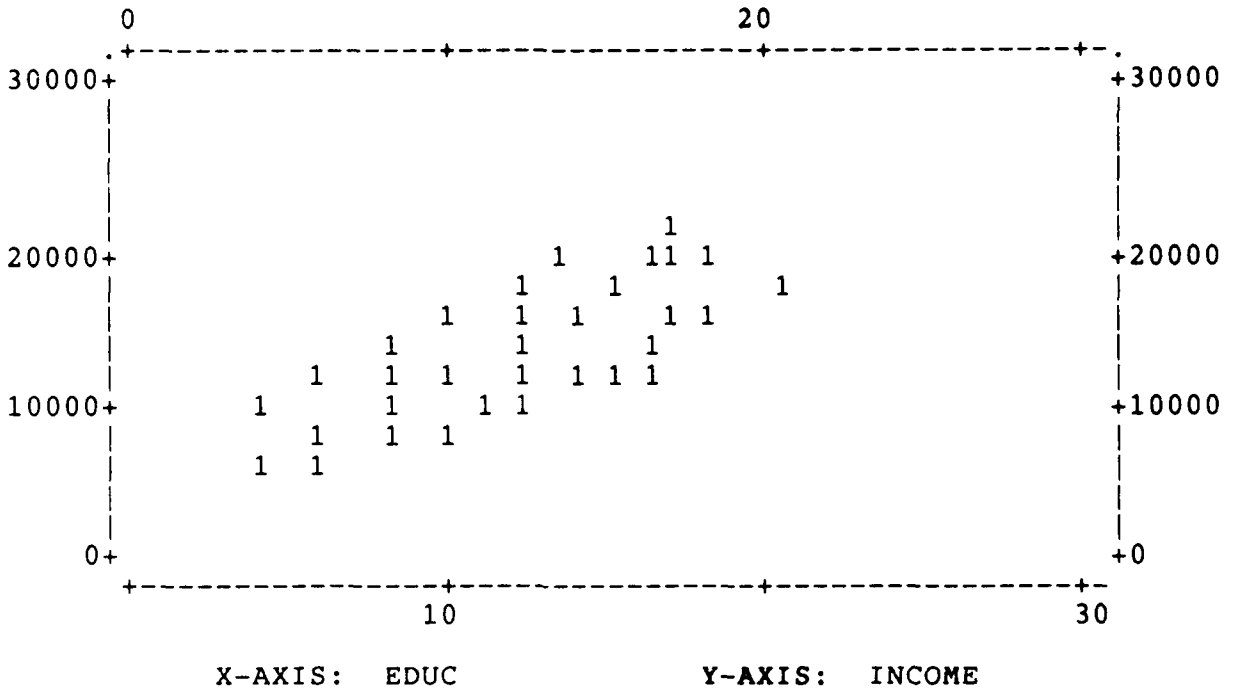


Figure 2. Scattergram of Income Versus Education

4) Linear Correlation Coefficients:

	AGE	EDUC	INCOME	NC
AGE	1.000	-0.201	-0.229	0.197
EDUC	-0.201	1.000	0.751	0.221
INCOME	-0.229	0.751	1.000	0.063
NC	0.197	0.221	0.063	1.000

5) Coefficients of Determination:

	AGE	EDUC	INCOME	NC
AGE	1.000	0.040	0.053	0.039
EDUC	0.040	1.000	0.564	0.049
INCOME	0.053	0.564	1.000	0.004
NC	0.039	0.049	0.004	1.000

The data are artificial except for education and income[5]. AGE was entered whimsically, and number of children was determined by random generation of numbers between 1 and 4. The correlations reflect the character and relation of the variables.

6) Random Sample - Descriptive Statistics

Responding to the prompts as shown,

WHAT FILE:

STATISTICS

WHAT VARIABLE:

AGE

ENTER THE NUMBER OF CASES DESIRED FOR THE SAMPLE (N>4): 20

ENTER (CASE) NUMBER OF LOWER LIMIT OF THE SAMPLE RANGE: 1

ENTER (CASE) NUMBER OF UPPER LIMIT OF THE SAMPLE RANGE: 32

twenty cases are selected randomly (with replacement) and the descriptive statistics are calculated (figure 6).

The option list presented when choosing the STATISTICS option of the FM main menu is generated by the FM routine ^DIX, the first part of which is shown in figure 4 (with the programmer additions underlined). As you add an option, making the numbers in line DIX+4 correspond to the total of options available stores your new option in the global ^DOPT and subsequently presents it to the user. Great care should be exercised in changing any FM routine that you do not rename. (The ^DIX routine is the link to the FM system, so the name ^DIX must be kept even though altered.) The original ^DIXC routine was completely rewritten and renamed ^T2STAT1 (figure 3). In addition, the routines shown in figures 5, 7, and 8 were added to the system.

In my opinion, there are two remarkable lines of code in the descriptive statistics (^T2STAT1) routine. They are line L in the SQR (square root) subroutine and line S2 in the SORT subroutine. Line L is Heron's Algorithm, which is derived from the Newton-Raphson's iteration formula[6]

$$X_{i+1}=X_i-f(xi)/f'(x_i).$$

This standard piece of code is easily my favorite because of its simplicity yet incredible power. Equally pleasing is the Shell sort[7] represented in S2.

It is remarkably efficient[8] and a fine example of exploiting the conditional transfer in iteration. This sort subroutine divides a list into two parts, compares the first element of the first half with the first element of the second half, and switches them if the first is greater than the second. After ordering the halves in this manner, the list is divided into quarters and the process is repeated. The final pass compares adjacent elements. This subroutine has been coded in BASIC, FORTRAN, PASCAL, and C, but none surpasses the elegance of the MUMPS version.

There are two coding styles evident in the ^T2STAT routines - the linear algebra approach, reflecting the treatment of relationships as sets of ordered pairs, and MUMPS global oriented. For example, contrast finding the maximum of a list by the MINMAX subroutine (or fetching it from ^DOSV(0,"TTB2:" 0,1, "H")) to simply writing it out after SORT.

To follow the indexing in obtaining the frequency counts for the percentile rankings, I suggest the reader comment out the kills (place a ";" in front of K) and look at the subscripts of the variables involved. This is a non-trivial coding structure and is probably unnecessarily complicated. The percentiles themselves are derived using linear interpolation; consequently, they are not actual variables, only location points.

^T2STAT3, in selecting a random sample and displaying descriptive statistics, departs entirely from the use of the ^DOSV statistical global and utilizes the user created global ^DIZ. To follow the code one must be aware that the routine ^DICRW places in the symbol table the variable % equal to the number of entries in the file, and the variable Y equal to "FILE#, FILENAME" - in our case Y="1000,STATISTICS". This routine has not been user "proofed" but was included to demonstrate (1) the generality of ^T2STAT1 and (2) the ease of application development using FM routines as utilities.

```
T2STAT1  ;DRH/NHRC ; 1/1/87 ; DESCRIPTIVE STATISTICS
         ;;17.07;VA FILEMAN;
D        D DESC G DESCX
         ;THE FOLLOWING SUBROUTINES ASSUME DATA SORTED IN ASCENDING ORDER OF
MAGNITUDE
MEAN     F J=1:1:SZT S SMX=0 D L0
         Q
L0       F K=1:1:XN(J) S SMX=SMX+Y(J,K)
         S M(J)=SMX/XN(J)
         Q
         ;
MEDIAN   I XN(J)#2 F J=1:1:SZT S MID(J)=Y(J,XN(J)+1/2)
         E  F J=1:1:SZT S MID(J)=(Y(J,XN(J)/2)+Y(J,XN(J)/2+1))/2
         Q
         ;
MODE     F J=1:1:SZT D L1
         Q
L1       F L=1:1:XN(J) S FRQ(J,L)=0,S(J,L)=Y(J,L) F K=1:1:XN(J) I S(J,L)=
Y(J,K) S FRQ(J,L)=FRQ(J,L)+1
         S MFQ(J)=1 F L=1:1:XN(J) S:FRQ(J,L)>MFQ(J) MFQ(J)=FRQ(J,L),MQ(J)=
"Y("_J_","_L_")"
         I MFQ(J)>1 S MD(J)=@MQ(J)_" OCCURS "_MFQ(J)_" TIMES "
         I '$D(MD(J)) S MD(J)="NO MULTIPLE OCCURRENCES"
         K S Q
         ;
PTILE    F J=1:1:SZT D L2
         Q
L2       S K=0,F(J,0)=1,(PR(J,0),PS(J,0))="" F L=1:1 S K=K+F(J,L-1)
Q:'$D(FRQ(J,K))  S F(J,L)=FRQ(J,K),PS(J,L)=Y(J,K)
         S CFQ(J,0)=0 F L=1:1 Q:'$D(F(J,L))  S CFQ(J,L)=CFQ(J,L-1)+F(J,L),
PR(J,L)=$E((CFQ(J,L)/XN(J))*100,1,6)
         F L=1:1 I PR(J,L)>25!(PR(J,L)=25) S PR2(J)=PR(J,L),PS2(J)=PS(J,L),
PR1(J)=PR(J,L-1),PS1(J)=PS(J,L-1) Q
         F L=1:1 I PR(J,L)>75!(PR(J,L)=75) S PR4(J)=PR(J,L),PS4(J)=PS(J,L),
PR3(J)=PR(J,L-1),PS3(J)=PS(J,L-1) Q
         S PS25(J)=(25-PR1(J))*(PS2(J)-PS1(J))/(PR2(J)-PR1(J))+PS1(J)
         S PS75(J)=(75-PR3(J))*(PS4(J)-PS3(J))/(PR4(J)-PR3(J))+PS3(J)
         K FRQ(J),CFQ(J),F(J),PR(J),PS(J) Q
         ;
MINMAX   F J=1:1:SZT S H(J)=Y(J,XN(J)),L(J)=Y(J,XN-XN(J)+1)
         Q
         ;
STDEV    F J=1:1:SZT S (SMX,SMXSQ)=0 D L3
         Q
L3       F K=1:1:XN(J) S SMX=SMX+Y(J,K),SMXSQ=SMXSQ+(Y(J,K)*Y(J,K))
         S VT(J)=SMX*SMX/XN(J),V(J)=(SMXSQ-VT(J))/(XN(J)-1)
         S X=V(J) D SQR S SD(J)=Y Q
         ;
SEM      F J=1:1:SZT S X=XN(J) D SQR S SXN(J)=Y,SEM(J)=SD(J)/SXN(J)
         Q
         ;
```

Figure 3. Descriptive Statistics Routine

```
SORT      F J=1:1:SZT D S1
S1        S G=XN(J)\2
S2        F I=G+1:1:XN(J) S L=I-G I (Y(J,L)>Y(J,L+G)) S T=Y(J,L),Y(J,L)=
Y(J,L+G),Y(J,L+G)=T G S2
S G=G\2 Q:G'>0  G S2
          Q
          ;
SQR       S Y=0 Q:X'>0  S Y=1+X/2
L         S T=Y,Y=X/T+T/2 G L:Y<T
          K T Q
          ;
DLCOR     D ^%ZIS S DJ=IO(0),U="^"
          I '$D(^DOSV(0,IO(0),0,1,"N")) Q
          F SZT=1:1 Q:'$D(^DOSV(0,DJ,0,SZT,"S"))  S XDN(SZT)=$E($P(^DOSV(0,
IO(0),"F",SZT),U,3),1,8),FN(SZT)=$E($P(^DOSV(0,IO(0),"F",SZT),U,1),1,9),
FP(SZT)=$E($P(^DOSV(0,IO(0),"F",SZT),U,2),1,2)+1
          S SZT=SZT-1,XN=$E($P(^DIZ(FN(1),0),U,4),1,5)
          F J=1:1:SZT F K=1:1:XN S X(J,K)=$E($P(^DIZ(FN(J),K,0),U,FP(J)),1,15)
          Q
DESC      ;CALCULATE DESCRIPTIVE STATISTICS.
          D DLCOR Q:'$D(^DOSV(0,IO(0),"F"))  F I=1:1:SZT I $D(^DOSV(0,DJ,0,I,
"Q")) S (^DOSV(0,DJ,0,I,"V"),X)=^("Q")-((^("S")*^("S"))/^("N"))/(^("N")-1) D
SQR S ^("D")=Y
          I '$D(^DOSV(0,IO(0),0,1,"S"))!('$D(^DOSV(0,IO(0),0,1,"N"))) W !!,"YOU
MUST PRECEDE VARIABLE NAME IN ""PRINT FIELD:"" OPTION WITH ""+"" OR ""#""" Q
          F J=1:1:SZT S XN(J)=XN F K=1:1:XN S Y(J,K)=X(J,K) I X(J,K)']$C(1) S
XN(J)=XN(J)-1
          K X
          D SORT,MEAN,MEDIAN,MODE,PTILE,MINMAX,STDEV,SEM
          Q
          ;
DESCX     ;PRINT DESCRIPTIVE STATISTICS
          I '$D(^DOSV(0,IO(0),0,1,"S"))!('$D(^DOSV(0,IO(0),0,1,"N"))) Q
          I '$D(^DOSV(0,IO(0),"F")) W *7,!!,?5,"NO STATISTICAL DATA ENTERED" Q
PRINT     K DHDR S DHDR="77CUST",DHDR(1)="DESCRIPTIVE STATISTICS" D DHDR^DIX
          W "I. MEASURES OF CENTRAL TENDENCY :",!!
          W ?1,"VARIABLE",?15,"N",?26,"MEAN",?40,"MEDIAN",?56,"MODE"
          F I=1:1:SZT W !,$J(XDN(I),8),?8 W
$J(XN(I),9),?20,$J(M(I),11,3),?35,$J(MID(I),11,3),?46,$J(MD(I),30)
          W !!,"II. MEASURES OF DISPERSION :",!!,?1,"A. ACROSS THE RANGE OF
DATA :",?51,"25%TILE",?66,"75%TILE"
          W !,?1,"VARIABLE",?15,"N",?24,"MINIMUM",?38,"MAXIMUM",?52,"SCORE",
?67,"SCORE"
          F I=1:1:SZT W!,$J(XDN(I),8),?10,$J(XN(I),6),$J(L(I),15,3),$J(H(I),
15,3), ?46,$J(PS25(I),15,2),$J(PS75(I),15,2)
          W !!," ","B. ABOUT THE MEAN :",?38,"STANDARD",?50,"STANDARD
ERROR",!,?1,"VARIABLE",?15,"N",?24,"VARIANCE",?38,"DEVIATION",?52,"OF THE
MEAN"
          F I=1:1:SZT W !,$J(XDN(I),8),?10,$J(XN(I),6),$J(V(I),15,3),$J(SD(I),
15,3),$J(SEM(I),15,3)
KL        I IO(0)'=IO S X=IO X ^DD("FUNC",7,1)
           U IO(0) K CFQ,DHDR,SD,SEM,SMXSQ,SXN,Y,V,VT,DJ,DTIME,ERR,F,FN,FP,FRQ,
G,H,I,J,K,L,M,N,NJ,MD,MFQ,MID,MQ,POP,PR,PR1,PR2,PR3,PR4,PS,PSMX,PS1,PS2,PS3,PS
4,PS25,PS75,R,SMX,SQMX,SQR,SUMPX,SUMSQ,S,SN,SZ,SZT,X,XDN,XN Q
```

Figure 3. Descriptive Statistics Routine (Cont'd)

```
DIX       ;GFT/SF ; 28FEB1986 6:08 PM
          ;;17.07;VA FILEMAN;
          D DT^DICRW
          S DIK="^DOPT(""DIX"",",IO(0)=$I
          G F:$D(^DOPT("DIX",6)) S ^(0)="STATISTICAL ROUTINE^1N^" F I=1:1:6 S
^DOPT("DIX",I,0)=$E($T(F+I),4,99)
          D IXALL^DIK
F         S DIC=DIK,DIC(0)="AEQZ" D ^DIC Q:Y<0  D @($P(Y(0),U,2,3)) W !! G DIX
          ;;DESCRIPTIVE STATISTICS^D^T2STAT1
          ;;SCATTERGRAM^^DIG
          ;;HISTOGRAM^^DIH
          ;;ESTIMATED LINEAR CORRELATION COEFFICIENTS^C^T2STAT2
          ;;COEFFICIENTS OF DETERMINATION^D^T2STAT2
          ;;RANDOM SAMPLE - DESCRIPTIVE STATISTICS^^T2STAT3
                                    .
                                    .
                                    .
```

Figure 4. Menu Program DIX

```
T2STAT2  ;DRH/NHRC ; 1/1/87 ; CORRELATION AND COEFFICIENT MATRIXES
         ;;17.07;VA FILEMAN;
C        D CORR G CORRX
D        D CORR G COEF
         ;
CORR     ;CALCULATE THE CORRELATION MATRIX
         D DLCOR^T2STAT1
         K ERR I $N(^DOSV(0,IO(0),0,1))'>0 W !!,"     AT LEAST TWO VARIABLES
MUST BE DEFINED     " S ERR=1 Q
         I '$D(^DOSV(0,IO(0),0,1,"H")) W !!," YOU MUST PRECEDE VARIABLE NAME
IN THE ""PRINT FIELD:"" OPTION WITH ""#""" S ERR=1 Q
         F J=1:1:SZT I ^DOSV(0,DJ,0,J,"H")=^("L") W !,"CAN'T COMPUTE
CORRELATION MATRIX-",XDN(J)," IS SINGLE-VALUED" S ERR=1 Q
         F J=1:1:SZT S XN(J)=XN F K=1:1:XN S Y(J,K)=X(J,K) I X(J,K)']$C(1) S
XN(J)=XN(J)-1
         D SORT^T2STAT1
         D MEAN^T2STAT1
         F J=1:1:SZT F K=1:1:XN(J) S SMX(J,K)=X(J,K)-M(J)
         S J=0 F NJ=SZT-1:-1:1 S J=J+1 F I=1:1:NJ S SUMPX(J,I+J)=0 F K=1:1:
XN(J) S PSMX(J,I+J,K)=SMX(J,K)*SMX(I+J,K),SUMPX(J,I+J)=SUMPX(J,I+J)+PSMX(J,
I+J,K)
         F J=1:1:SZT S SUMSQ(J)=0 F K=1:1:XN(J) S SQMX(J,K)=SMX(J,K)*SMX(J,K),
SUMSQ(J)=SUMSQ(J)+SQMX(J,K)
         F J=1:1:SZT S X=SUMSQ(J) D SQR^T2STAT1 S SQR(J)=Y
         S J=0 F NJ=SZT-1:-1:1 S J=J+1 F I=1:1:NJ S R(J,I+J)=SUMPX(J,I+J)/
(SQR(J)*SQR(I+J)),R(I+J,J)=R(J,I+J),RSQ(J,I+J)=R(J,I+J)*R(J,I+J),RSQ(I+J,J)=
R(I+J,J)*R(I+J,J)
         Q
CORRX    ;PRINT THE CORRELATION MATRIX
         G:$D(ERR) KL K DHDR S DHDR="72TSU",DHDR(1)="",DHDR(2)="" D DHDR^DIX
         W !!!,"CORRELATION MATRIX:",!!
         F L=1:1:SZT S R(L,L)=1
         W ?10 F J=1:1:SZT W $J(XDN(J),10)
         F J=1:1:SZT W !!,$J(XDN(J),10) F K=1:1:SZT W $J(R(J,K),10,3)
         G KL
COEF     ;PRINT THE COEFFICIENTS OF DETERMINATION
         G:$D(ERR) KL K DHDR S DHDR="72TSU",DHDR(1)="",DHDR(2)="" D DHDR^DIX
         W !!!,"COEFFICIENTS OF DETERMINATION:",!!
         F L=1:1:SZT S RSQ(L,L)=1
         W ?10 F J=1:1:SZT W $J(XDN(J),10)
         F J=1:1:SZT W !!,$J(XDN(J),10) F K=1:1:SZT W $J(RSQ(J,K),10,3)
KL       I IO(0)'=IO S X=IO X ^DD("FUNC",7,1)
         U IO(0) K CFQ,DHDR,RSQ,DJ,DTIME,ERR,F,FN,FP,FRQ,G,I,J,K,L,M,N,NJ,MD,
MFQ,MID,MQ,POP,PR,PR1,PR2,PR3,PR4,PS,PSMX,PS1,PS2,PS3,PS4,PS25,PS75,R,SMX,
SQMX,SQR,SUMPX,SUMSQ,S,SZ,SZT,X,XDN,XN,Y Q
```

Figure 7. Routine to Calculate and Print Correlation and Coefficient Matrixes

I. MEASURES OF CENTRAL TENDENCY:

VARIABLE	N	MEAN	MEDIAN	MODE
AGE	20	41.450	39.000	52 OCCURS 4 TIMES

II. MEASURES OF DISPERSION:

A. ACROSS THE RANGE OF DATA:

VARIABLE	N	MINIMUM	MAXIMUM	25%TILE SCORE	75%TILE SCORE
AGE	20	20.000	76.000	28.33	49.50

B. ABOUT THE MEAN:

VARIABLE	N	VARIANCE	STANDARD DEVIATION	STANDARD ERROR OF THE MEAN
AGE	20	207.945	14.420	3.224

Figure 6. Descriptive Statistics for a Random Sample

```
T2STAT3  ;DRH/NHRC; 1/1/87 ;SELECTS RANDOM SAMPLE AND DOES DESCRIPTIVE
STATISTICS
         I '$D(^YRN(1)) S ^YRN(1)=813640269944916720
         D ^%ZIS S U="^",SZT=1,S=0,Z=1
         R "WHAT FILE: ",!,X Q:X=""  S DIC="^DIC(",DIC(0)="QE" D DIC+3^DICRW
         R !,"WHAT VARIABLE: ",!,XDN(Z)
         S XN=%,C="",FN(Z)=$E(Y,1,4),FP(Z)=$N(^DD(FN(Z),"B",XDN(Z),C))+1
RAND     D ^T2STAT4 F K=1:1:SN S Y(Z,K)=$E($P(^DIZ(FN(Z),Y(K),0),U,FP(Z)),1,
15)
         S (XN,XN(Z))=SN D SORT^T2STAT1,MEAN^T2STAT1,MEDIAN^T2STAT1,
MODE^T2STAT1,PTILE^T2STAT1,MINMAX^T2STAT1,STDEV^T2STAT1,SEM^T2STAT1
         D PRINT^T2STAT1
         I IO(0)'=IO S X=IO X ^DD("FUNC",7,1) Q
```

Figure 7. Program for Random Sample

```
T2STAT4  ; DRH,NHRC; 1/1/87;  GENERATES SN RANDOM NUMBERS  X1<=Y(i)<=X2
0<=i<=SN
         ; NOTE: A RANDOM 19 DIGIT INTEGER MUST BE SEEDED IN ^YRN(1)
ST1      S FLAG=S R !,"ENTER THE NUMBER OF CASES DESIRED FOR THE SAMPLE (N>1)
: ",SN
         R !,"ENTER (CASE) NUMBER OF LOWER LIMIT OF THE SAMPLE RANGE :",X1
         R !,"ENTER (CASE) NUMBER OF UPPER LIMIT OF THE SAMPLE RANGE :",X2
ST2      S B=0,C=0,D=1
         F J=10,100,1000,10000,100000 S B=B+1 I X2/J<1 S EXT=B Q
         I X2-X1=1 S EXT=6,D=10000
         I X2'>1 S D=100000,EXT=6
RNG      S Y1(1)=^YRN(1),Y2(1)=$E($E(^YRN(1),7,11),1,EXT)/D,K=262147
         F %=2:1:SN+1 S Y1(%)=$E(Y1(%-1)*K,1,19),Y2(%-1)=$E($E(Y1(%),7,11),
1,EXT)/D
         I X1=1 F J1=1:1:SN S Y2(J1)=Y2(J1)+1
SEC      F J3=1:1:SN I (Y2(J3)'>X2)&(Y2(J3)'<X1) S C=C+1,Y(C)=Y2(J3)
         S ^YRN(1)=$E(Y1(SN+1),1,19)
         I C<SN G RNG
         I FLAG W !!,SN," RANDOM NUMBER"_$S(SN>1:"S",SN=1:"",1:""),"  FROM THE
RANGE ",X1," TO ",X2," "_$S(SN>1:"ARE",SN=1:"IS",1:""),"  : ",! F J=1:1:SN W
!,?19,J,".  ",Y(J)
         I FLAG K FLAG,SN,Y,S
END      K FLAG,Y1,Y2,B,C,D,L,M,K,X1,X2,J,J1,J2,J3,%,EXT Q
```

(NOTE: Set S=1 for this routine to generate random numbers in any range).

Figure 4. Random Number Generator

^T2STAT4, the random number generator, has been tested for integers and "passes" Pearson's chi square statistic for various degrees of freedom at p=.95. It is set up to generate random numbers in any range, i.e. between 0 and 1, .7 and .85 etc., but has not been evaluated for all ranges. If one sets S=1 and then commands D ^T2STAT4, the routine functions as a straight random number generator and prints the numbers requested. The variable ^YRN(1) must be set to a 19-digit random number prior to using the routine; thereafter seeding is automatic. Anyone with a compelling need to handle a large database with these routines can rename local variables, like Y(J,K) and X(J,K), global variables. The computing time will go up appreciably because of disk reads, however. Another way to process large files is to transfer data making maximum disk reads using $S to measure memory in the partition. This method fails on our present ISM M/VMS system as the partition tables are not being entirely cleared. This results in $S returning erratic values, thus making it impossible to control disk reads as a function of $S.

Discussion

NOHIMS must provide for the processing, storage, and archiving of information pertaining to approximately 200,000 men and women at approximately 162 separate locations. The critical strength in the choice of FM is that it can be used as a relational database management system (RDBMS). The inclusion in Version 17.0 of the colon (":") syntax in the search, sort, and print field specifications allows the user to specify fields in files that point to the file of origin (the backward pointer). There now exist three relational navigation routes employing the ":" syntax: (1) specifying fields in files pointed to by a pointer valued field in the file of origin; (2) using the ":" as an extended pointer to move from the original file to any other file; and (3) specifying fields in a file that have a field pointing to the file of origin, as mentioned[4]. With these improvements, FM is more symmetric and can be viewed as a data dictionary driven RDBMS.

To discover what relationships exist in the database, one has only to choose the PRINT FILES option of the FM menu and respond "?" to the prompt "OUTPUT FROM WHAT FILE:" He is immediately graced with a ternary relationship "FILE LIST" with three domains: file number, file name, and number of entries in the files. In the case at hand, the file STATISTICS has file number 1000 and 32 sets. "SETS OF WHAT?" is answered by choosing LIST FILE ATTRIBUTES and listing the file attributes conveniently stored in the data dictionary. Any of these items can be altered or deleted without diminishing the usefulness of present programs or future access.

Using the mathematic model of a relationship R defined on n sets, we can look at the global ^DIZ as having the relationship STATISTICS defined on the domains NAME, AGE, EDUC, INCOME, and NC. There are thirty-two sets of 5 tuples, each with an element from one of the five domains. This view of the data as an array amenable to set operations such as intersect, project, select, and join makes possible a data manipulation methodology that allows the construction of data arrays free from the problems of ordering, indexing, and access path dependence[2,3]. (Note the device number "TTB2:" in ^DOSV. This is an example of dependence that results in the need to recreate ^DOSV if

one changes device numbers.) If, in your mind's eye, you can visualize Venn diagrams consisting of files with records as the member sets, the algebra of joins, intersects, and complements, suggests itself.

The operational economy of a database manipulated algebraically and treating relationships as sets of ordered pairs is without question. The expression Y(J,K) used here carries the following information: (1) Y is a general symbol for the relationship name, i.e., STATISTICS (J,K). In conventional terminology, Y is the file; (2) J is the general symbol for the domain - conventionally the variable name - in our example, AGE, etc.; (3) K is the general symbol for the set number - conventionally the record number of a file (K is also the record primary key. The indexing in the ^T2STAT routines assumes the user has not redefined the field "number" (field number .001) in FM). Thus, Y(J,K) translates to STATISTICS(AGE, RECORDNUMBER)--a relationship Y on domain J, range K.

There is little question of accuracy in matrix operations on the VAX 8200 using M/VX which has 18 decimal digit accuracy with 64 bit precision. Non-integers are stored base 10 (versus the traditional floating point) ensuring full 64 bit representation. This is slower, but accurate. Double precision arithmetic employed at critical points in matrix operations helps bound errors at acceptable levels.

Conclusion

The Veteran's Administration File Manager, with the addition of relational syntax, provides a powerful relational database management environment. File information, data description (and documentation), and data structuring are easily within the grasp of the most casual user because it is data dictionary driven. Creating programs to analyze data is simplified with a relational view of the database. The myriad of problems that make data inaccessible are minimized in a relational database context.

The objective, of course, is to have a data structure independent of application programs and easy to access. The integration of data that are consistent and free from redundancy from 162 sources is most likely to be achieved in a relational context. Using FM as a RDBMS serves all levels of demand equally well. The industrial hygienist can explore the database as

easily as the statistician. The use of statistical packages such as SPSS is facilitated and archiving becomes a meaningful activity instead of a burial ceremony.

References

1. Pugh, W.M., Beck, D.D., Ramsey-Klee, D.M. An Overview of the Navy Occupational Health Information Monitoring System (NOHIMS), Naval Health Research Center, P. O. Box 85122, San Diego, CA 92138 (1984).

2. Codd, E.F., A Relational Model for Large Shared Data Banks, Comm. ACM Vol. 13(6):33-64 (1970).

3. O'Kane, K.C.,. Design for a Relational Data Base System in MUMPS, MUG Quarterly Vol. XV(2):33-36 (1986).

4. Timson, G. VA Fileman Version 17 Release Notes (1986), personal communication.

5. Lewis-Beck, M.S. Applied Regression, Sage Publications, Beverly Hills, p. 16 (1980).

6. Korn, G.A., Korn, T.M. Mathematical Handbook for Scientists and Engineers, San Francisco: McGraw-Hill Book Company, 20.2-2 (1968).

7. Shell, D.L. CACM 2, pp. 30-32 (July 1959).

8. Knuth, D. D. The Art of Computer Programming, Menlo Park, California, Addison-Wesley Publishing Company, Vol. 3:84-95 (1973).

REPORT DOCUMENTATION PAGE

A186097

1b RESTRICTIVE MARKINGS
None

3 DISTRIBUTION AVAILABILITY OF REPORT
Approved for public release; distribution unlimited.

5 MONITORING ORGANIZATION REPORT NUMBER(S)

6b OFFICE SYMBOL (If applicable)

7a NAME OF MONITORING ORGANIZATION
Commander, Naval Medical Command

7b ADDRESS (City, State and ZIP Code)
Department of the Navy
Washington, D.C. 20372

8b OFFICE SYMBOL (If applicable)

9 PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER

10 SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO	PROJECT NO	TASK NO	WORK UNIT ACCESSION NO
63706N	M0095	.005	DN246555

11 (INCLUDE SECURITY CLASSIFICATION) EPIDEMIOLOGIC STATISTICS USING THE VETERANS ADMINISTRATION FILE MANAGER AS A RELATIONAL DATABASE MANAGEMENT SYSTEM

12 PERSONAL AUTHOR(S)

13a TYPE OF REPORT

13b TIME COVERED FROM ____ TO ____

14 DATE OF REPORT (Year, Month, Day)
87 March 16

15 PAGE COUNT
1

16 SUPPLEMENTARY NOTATION

17 COSATI CODES

FIELD	GROUP	SUB-GROUP

18 SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
Epidemiologic Statistics, Relational Database, Veterans Administration File Manager

19 ABSTRACT (Continue on reverse if necessary and identify by block number)

Epidemiologic statistics programs underscore the efficacy of a relational [illegible] which not only [illegible] manipulation allow sufficient multi-[illegible] statistics. The Veterans Administration File Manager is viewed as a [illegible] management system.

20 DISTRIBUTION/AVAILABILITY OF ABSTRACT
☒ UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT ☐ DTIC USERS

21 ABSTRACT SECURITY CLASSIFICATION
Unclassified

22a NAME OF RESPONSIBLE INDIVIDUAL

22b TELEPHONE (Include Area Code)
(619)225-2071

22c OFFICE SYMBOL
Code 2

DD FORM 1473 83 APR edition may be used until exhausted
All other editions are obsolete

SECURITY CLASSIFICATION OF THIS PAGE

GPO 1988-607-047

END

DATE

FILMED

DEC.

1987